CLAIMS

1. A method for automatically filtering a corpus of documents containing textual and non-textual information of a natural language, the method being characterized in that it comprises the steps of:

- dividing the corpus of documents into appropriate portions;
- determining for each portion of the corpus of documents a regularity value (V_R) measuring the conformity of the portion with respect to character sequences probabilities predetermined for said language;
- comparing each regularity value with a threshold value (V_T) to decide whether the conformity is sufficient; and
- rejecting any portion of the corpus of documents whose conformity is not sufficient.

2. Method according to Claim 1, wherein said character sequences probabilities is derived from a statistical model representative of said language.

3. Method according to Claim 2, wherein for each portion of the corpus of documents, said regularity value (V_R) is based on a computed perplexity of the portion with respect to said statistical model.

4. Method according to Claim 2, wherein said statistical model is previously elaborated from a reference document determined as conforming with the rules of said language.

5. Method according to Claim 2, wherein said statistical model is being determined according to N-gram statistics.

6. Method according to Claim 2, wherein said statistical model is a character-based N-gram model.

7. Method according to Claim 2, wherein said statistical model is initially used to filter a first corpus segment of a predetermined size to provide a first filtered segment of the corpus of documents, said first filtered segment serving as a basis for computing a more accurate statistical model which is to be used to filter the rest of the corpus of documents.

8. Method according to Claim 1, wherein said threshold value (V_T) is determined by executing the following steps of:

- defining a test corpus as a subset of the corpus of documents to be filtered;
- manually cleaning said test corpus so as to obtain a cleaned test corpus which is representative of the type of textual information that is considered as being sufficiently in conformity with the language rules and a rejected test corpus that is the complement of said cleaned test corpus;
- computing a perplexity value for each of said cleaned and rejected test corpora with regard to said statistical model; and
- setting the threshold value searched between the perplexity values computed.

9. Method according to Claim 1, wherein said portions comprise lines, paragraphs, and whole documents – whose size is determined as a function of the overall size of the corpus of documents or as a function of the nature of the documents contained in the corpus of documents or both, so as to obtain a granularity desired for the filtering.

10. An apparatus for automatically filtering a corpus of documents containing textual and non-textual information of a natural language, the apparatus being characterized in that it comprises:

- means for dividing the corpus of documents into appropriate portions;
- means for determining for each portion of the corpus of documents a regularity value measuring the conformity of the portion with respect to character sequences probabilities predetermined for said language;

- means for comparing each regularity value with a threshold value to decide whether the conformity is sufficient; and

- means for rejecting any portion of the corpus of documents whose conformity is not sufficient.

11. Apparatus according to Claim 10, wherein said character sequences probabilities are derived from a statistical model representative of said language.

12. Apparatus according to Claim 11, wherein for each portion of the corpus of documents, said regularity value (V_R) is based on a computed perplexity of the portion with respect to said statistical model.

13. Apparatus according to Claim 11, wherein said statistical model is previously elaborated from a reference document determined as conforming with the rules of said language.

14. Apparatus according to Claim 11, wherein said statistical model is being determined according to N-gram statistics.

15. Apparatus according to Claim 11, wherein said statistical model is a character-based N-gram model.

16. Apparatus according to Claim 11, wherein said statistical model is initially used to filter a first corpus segment of a predetermined size to provide a first filtered segment of the corpus of documents, said first filtered segment serving as a basis for computing a more accurate statistical model which is to be used to filter the rest of the corpus of documents.

17. Apparatus according to Claim 10, wherein said threshold value (V_T) is determined by executing the following steps of:

- defining a test corpus as a subset of the corpus of documents to be filtered;

- manually cleaning said test corpus so as to obtain a cleaned test corpus which is representative of the type of textual information that is considered as being sufficiently in conformity with the language rules and a rejected test corpus that is the complement of said cleaned test corpus;

- computing a perplexity value for each of said cleaned and rejected test corpora with regard to said statistical model; and

- setting the threshold value searched between the perplexity values computed.

18. Apparatus according to Claim 10, wherein said portions comprise lines, paragraphs, and whole documents – whose size is determined as a function of the overall size of the corpus of documents or as a function of the nature of the documents contained in the corpus of documents or both, so as to obtain a granularity desired for the filtering.

19. A computer system comprising an apparatus according to Claim 10.

20. A computer program comprising software code portions for performing a method according to Claim 1, when said computer program is loaded and executed by a computer system.

21. A computer-readable program storage medium which stores a program for executing a method for automatically filtering a corpus of documents containing textual and non-textual information of a natural language, the method being characterized in that it comprises the steps of:

- dividing the corpus of documents into appropriate portions;

- determining for each portion of the corpus of documents a regularity value (V_R) measuring the conformity of the portion with respect to character sequences probabilities predetermined for said language;

- comparing each regularity value with a threshold value (V_T) to decide whether the conformity is sufficient; and

- rejecting any portion of the corpus of documents whose conformity is not sufficient.

22. Computer-readable program storage medium according to Claim 21, wherein said character sequences probabilities is derived from a statistical model representative of said language.

23. Computer-readable program storage medium according to Claim 22, wherein for each portion of the corpus of documents, said regularity value (V_R) is based on a computed perplexity of the portion with respect to said statistical model.

24. Computer-readable program storage medium according to Claim 22, wherein said statistical model is previously elaborated from a reference document determined as conforming with the rules of said language.

25. Computer-readable program storage medium according to Claim 22, wherein said statistical model is being determined according to N-gram statistics.

26. Computer-readable program storage medium according to Claim 22, wherein said statistical model is a character-based N-gram model.

27. Computer-readable program storage medium according to Claim 22, wherein said statistical model is initially used to filter a first corpus segment of a predetermined size to provide a first filtered segment of the corpus of documents, said first filtered segment serving as a basis for computing a more accurate statistical model which is to be used to filter the rest of the corpus of documents.

28. Computer-readable program storage medium according to Claim 21, wherein said threshold value (V_T) is determined by executing the following steps of:

- defining a test corpus as a subset of the corpus of documents to be filtered;

- manually cleaning said test corpus so as to obtain a cleaned test corpus which is representative of the type of textual information that is considered as being sufficiently in conformity with the language rules and a rejected test corpus that is the complement of said cleaned test corpus;

- computing a perplexity value for each of said cleaned and rejected test corpora with regard to said statistical model; and

- setting the threshold value searched between the perplexity values computed.

29. Computer-readable program storage medium according to Claim 21, wherein said portions comprise lines, paragraphs, and whole documents – whose size is determined as a function of the overall size of the corpus of documents or as a function of the nature of the documents contained in the corpus of documents or both, so as to obtain a granularity desired for the filtering.